Constraint-Based Clustering Selection

Toon Van Craenendonck TOON.VANCRAENENDONCK@CS.KULEUVEN.BE
Department of Computer Science, KU Leuven, Belgium

Hendrik Blockeel HENDRIK.BLOCKEEL@CS.KULEUVEN.BE
Department of Computer Science, KU Leuven, Belgium

Keywords: constraint-based clustering, algorithm and hyperparameter selection

Abstract

Semi-supervised clustering methods incorporate a limited amount of supervision into the clustering process. Typically, this supervision is provided by the user in the form of pairwise constraints. Existing methods use such constraints in one of the following ways: they adapt their clustering procedure, their similarity metric, or both. All of these approaches operate within the scope of individual clustering algorithms. In contrast, we propose to use constraints to choose between clusterings generated by very different unsupervised clustering algorithms, run with different parameter settings. We empirically show that this simple approach often outperforms existing semi-supervised clustering methods.

1. Introduction

Clustering is one of the core tasks in data analysis (Jain, 2010). It is inherently subjective, as users may prefer very different clusterings of the same data (Caruana et al., 2006; von Luxburg et al., 2014). Semi-supervised clustering (Wagstaff et al., 2001; Xing et al., 2003) aims to deal with this subjectivity by allowing the user to specify background knowledge, often in the form of pairwise constraints that indicate whether two instances should be in the same cluster or not.

Traditional approaches to semi-supervised (or constraint-based) clustering use constraints in one of the following three ways. First, one can modify an existing clustering algorithm to take them into account. This approach is taken in COP-KMeans (Wagstaff et al., 2001), a modification of K-Means in which points are assigned to their closest cluster such that no constraint is violated. Second, one can learn a distance metric based on the constraints (Xing et al., 2003), after which the metric is used within a traditional clustering algorithm. Third, one can combine the above two approaches and develop so-called hybrid methods (Bilenko et al., 2004).

Our approach to constraint-based clustering is quite different from existing methods, and does not fit in any of these three categories. It is motivated by the well-known fact that different algorithms may produce very different clusterings of the same data (Estivill-Castro, 2002), and even within one algorithm, different parameter settings may yield different clusterings. This implies that selecting a clustering algorithm and tuning its parameter settings is crucial to obtain a good clustering. We propose to use constraints to solve these tasks, as discussed in the next section.

2. Constraint-Based Selection

In an entirely unsupervised setting, selecting and tuning a clustering algorithm is difficult. This is mainly due to the lack of a well-defined way to estimate the quality of clustering results (Estivill-Castro, 2002). We propose to use constraints for this purpose, and estimate the quality of a clustering as the number of constraints that it satisfies. This quality estimate allows us to do a grid search over unsupervised algorithms and their parameter settings, as described in Algorithm 1. We call this approach to semi-supervised clustering COBS (for Constraint-Based Selection). We assume that we are given a set of must-link constraints ML, where $(i, j) \in ML$ indicates that instances x_i and x_j should be in the same cluster. Similarly, we are

An expanded version of this work is submitted as a conference paper.

given a set of cannot-link constraints CL, where $(i, j) \in CL$ indicates that x_i and x_j should be in different clusters. We select the "best" solution from a set of clusterings as the one satisfying the largest number of constraints (in case of a tie, we select randomly from the involved clusterings).

Algorithm 1 Constraint-based selection (COBS)

Input: D: a dataset
ML: set of must-link constraints
CL: set of cannot-link constraints
Output: a clustering of D
1: Generate a set of clusterings C by varying the hyperparameters of several unsupervised clustering algorithms
2: **return** $\arg\max_{c \in C} \left(\sum_{(i,j) \in ML} \mathbb{I}[c[i]=c[j]] + \sum_{(i,j) \in CL} \mathbb{I}[c[i] \neq c[j]] \right)$

3. Experimental evaluation

We run COBS with K-means, DBSCAN and spectral clustering as unsupervised clustering algorithms in step one of Algorithm 1, as they are common representatives of different types of algorithms. We compare our approach to individual semi-supervised variants of these algorithms: MPCKMeans (Bilenko et al., 2004), COSC (Rangapuram & Hein, 2012), and FOSC-OpticsDend (Campello et al., 2013). We assume that the only external input to the clustering algorithms is in the form of pairwise constraints. Consequently, the number of clusters K has to be selected for both MPCKMeans and COSC. The three curves for MPCKMeans and COSC in Figure 1 correspond to three different ways of doing this. In COBS, the number of clusters is treated as any other hyperparameter.

We show experiments with six UCI classification datasets. The classes are assumed to represent the clusters of interest. To generate pairwise constraints we repeatedly select two random instances, and add a must-link constraint if they belong to the same class, and a cannot-link otherwise. We evaluate how well the returned clusters coincide with the known classes by computing the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985).

Figure 1 shows clustering performance for an increasing number of pairwise constraints. COBS is the only approach that consistently produces good clusterings for the first four datasets. None of the methods is able to produce good clusterings for glass and hepatitis, suggesting that the class labels do not indicate a natural grouping. The figure shows that *it is often better to use constraints to select and tune an unsupervised algorithm, than within a randomly chosen semi-supervised algorithm.*

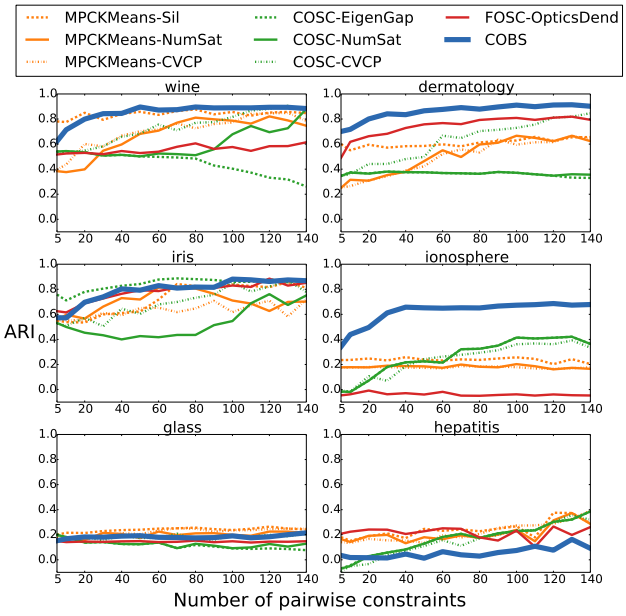

Figure 1. Performance of COBS vs. semi-supervised algorithms

4. Conclusion

Exploiting constraints has been the subject of substantial research, but all existing methods use them within the clustering process of individual algorithms. In contrast, we propose to use them to choose between clusterings generated by different unsupervised algorithms, ran with different parameter settings. We experimentally show that this strategy is superior to all the semi-supervised algorithms compared to, which themselves are state of the art and representative for a wide range of algorithms. For the majority of the datasets, it works as well as the best among them, and on average it performs much better.

Acknowledgments

This work is supported by the Agency for Innovation by Science and Technology in Flanders (IWT).

References

Bilenko, M., Basu, S., & Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. *Proc. of 21st International Conference on Machine Learning* (pp. 81–88).

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery*, *27*, 344–371.

Caruana, R., Elhawary, M., & Nguyen, N. (2006). Meta clustering. *Proc. of the International Conference on Data Mining*.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, *4*, 65–75.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, *31*, 651–666.

Rangapuram, S. S., & Hein, M. (2012). Constrained 1-spectral clustering. *Proc. of the 15th International Conference on Artificial Intelligence and Statistics*.

von Luxburg, U., Williamson, R. C., & Guyon, I. (2014). Clustering: Science or Art? *Workshop on Unsupervised Learning and Transfer Learning, JMLR Workshop and Conference Proceedings 27*.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proc. of the Eighteenth International Conference on Machine Learning* (pp. 577–584).

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2003). Distance metric learning, with application to clustering with side-information. *Advances in Neural Information Processing Systems 15* (pp. 505–512).